Discrepancies Between Score Trends from NAEP and State Tests: A Scale-Invariant Perspective

Discrepancies Between Score Trends from NAEP and State Tests: A Scale-Invariant Perspective

DRAFT

Discrepancies Between Score Trends from NAEP and State Tests:
A Scale-Invariant Perspective

State test score trends are widely interpreted as indicators of educational improvement. To validate these interpretations, state test score trends are often compared to trends on other tests such as the National Assessment of Educational Progress (NAEP). These comparisons raise serious technical and substantive concerns. Technically, the most commonly used trend statistics – for example, the change in the percent of proficient students – are misleading in the context of cross-test comparisons. Substantively, it may not be reasonable to expect that NAEP and state test score trends should be similar. This paper motivates then applies a "scale-invariant" framework for cross-test trend comparisons to compare "high-stakes" state test score trends from 2003 to 2005 to NAEP trends over the same period. Results show that state trends are significantly more positive than NAEP trends. The paper concludes with cautions against the positioning of trend discrepancies in a framework where only one trend is considered "true."

State test score trends are among the most prominent large-scale educational statistics. Positive trends are interpreted as an improvement in the education of students and as an increase in student learning. As states implement educational policies under the No Child Left Behind (NCLB) Act of 2001, trend statistics are also interpreted as evidence that these policies are or are not functioning as intended.

These high-stakes interpretations are largely supported by data from two sources. State testing programs include one or more test batteries and report aggregated state-level results by grade, subject, and year. The National Assessment of Educational Progress (NAEP) has provided state-level results at various intervals since 1990, most recently in 2003 and 2005.

Hereafter, this paper uses the shorthand "State" (with the capital "S") and "NAEP" to distinguish between these two data sources.

State testing programs and NAEP are both charged with reporting and describing what students in each state know and are able to do. Both State tests and NAEP provide a highly visible and interpretable summary measure, "Percent Proficient" or "Percent Proficient and Above," for each subject, grade and year and subgroup tested. NAEP has reported state-level results for the Spring of 2003 and 2005, for Reading and Mathematics, and for grades 4 and 8. Where State tests coincide with these subjects, grades, and years, it seems straightforward to assume that State and NAEP results will correspond.

These assumptions have led to a large number of reports, articles, and newspaper headlines. Reports that have taken advantage of the recent release of 2005 state NAEP results for State-NAEP comparisons include those by the Thomas B. Fordham Foundation (2005), the Education Trust (Hall & Kennedy, 2006), Education Week (2006), the Civil Rights Project at Harvard University (Lee, 2006), and Policy Analysis for California Education (Fuller, Gesicki, Kang, & Wright, 2006). A sampling of widely cited papers comparing State and NAEP results includes Linn, Graue, and Sanders (1990), Koretz and Barron (1998), Klein, Hamilton, McCaffrey, and Stecher (2000), and Linn, Baker and Betebenner (2002). Trend discrepancies are provocative. Given two tests that have the same name (e.g., Reading, Mathematics) and sample from the same population (e.g., 4th graders in California), juxtaposing discrepant trends frames them as contradictory.

In contrast, this paper considers the act of comparing State and NAEP results as the act of comparing the height of two children on pogo sticks. There are technical and substantive issues that arise in this comparison. Technically, it is impossible to compare the heights of these

children until the pogo sticks are removed. Substantively, once the pogo sticks are removed, it may be reasonable to ask, should we expect the heights of these children to be the same? The pogo sticks in this analogy represent the unsuitable properties of the trend statistics often used in State-NAEP comparisons. The significance of the children in this analogy is that they are not twins; neither are State tests and NAEP; and this fact should temper expectations about the similarity of results. To maintain this analogy, this paper will proceed by describing the problems with pogo sticks, comparing the children's heights without the pogo sticks, and listing some reasons why their heights might be expected to differ.

Technical Issues Arising in State-NAEP Comparisons

The perception and implementation of NCLB depend crucially on the concept of *proficiency* – a child is *left behind* if she or he is not proficient. In practice, this requires the dichotomization of test score scales into *proficient* and *not proficient* regions. This proficiency framework has led to the quantification of test score trends as *the change in the percent of proficient students*. The class of proficiency-based statistics can be generally described as Percent Above Cut (PAC) statistics, where the cut score in this case refers to the passing score that delineates proficiency. Trends in this framework are expressed as changes in PACs.

Both the pervasiveness of the PAC framework and the problems with the framework have been demonstrated on recent covers of this very publication. In the Fall 2006 issue of *Educational Measurement: Issues and Practice*, then Editor Steve Ferrara presented discrepant score trends, as measured by the change in the percent of proficient students, in a figure from the aforementioned Policy Analysis for California Education report (Fuller, Gesicki, Kang, & Wright, 2006). In an earlier, Spring 2005 issue, Ferrara covered a figure from Paul Holland's

2002 paper. This figure shows the dramatic dependency of the change-in-gap statistic on the selection of the cut score.

The choice of cover visuals is interesting in that one raises technical concerns about the other. The take-home message from Holland's (2002) paper is that PAC-based gap trends are easily sign-reversible under different selections of cut scores. For example, the achievement gap between two groups of students may be increasing in terms of the percent of proficient students while decreasing in terms of the percent of "basic" students. Though Holland's paper focused on gap trends as opposed to score trends, the extensions are straightforward. Holland's paper is under-cited given the popularity of PAC-based trend statistics, and this section briefly rephrases Holland's cautions on changes-in-gaps in the context of trends and trend comparisons. To return to the previous analogy, this discussion concerns the problems with pogo sticks.

The interpretive problems with PAC-based statistics may be simply explained by their interaction with unimodal distributions. If a unimodal distribution of test scores shifts in the positive direction, the rate at which examinees cross a cut score will not be constant. As the mode of the distribution approaches the cut score, more and more examinees will cross in equal units of time. After the mode of the distribution passes the cut score, fewer and fewer examinees will cross in equal units of time. If the cut score were different under this model, the trend would be different. In this sense, PAC-based trends may be described as *pliable* under the choice of cut score.

Two test score distributions may be represented by Cumulative Distribution Functions (CDFs) that return the proportion of examinees at or below a cut score. Succinctly, a CDF, $F(x)$ takes a score x and returns a proportion p of examinees at or below that score. Figure 1 shows two normally distributed CDFs from two time points on an arbitrary score scale. The

distributions have different means and standard deviations and are designated as $F_1(x)$ and $F_2(x)$. The unimodal nature of the distributions can be observed by noting the steep rise in the CDFs towards their modes where they accumulate the most examinees.

Insert Figure 1 About Here

Holland (2002) observes that the PAC-based trend can be viewed on CDFs as the vertical slice between the two CDFs. If $F_1(x)$ returns p_1, the percent of students at or below a proficiency cut score x, and $F_2(x)$ returns p_2, the percent at or below that same cut score, the change in the percent *above* a cut score is simply $(1 - p_2) - (1 - p_1)$ or $F_1(x) - F_2(x)$. The vertical difference between these two CDFs is displayed around the x-axis as a dotted line. These are the PAC-based trends for cut scores at those respective locations on the score scale. The figure shows that a different selection of cut score will change and even reverse the sign of the PAC-based trend and resulting trend interpretations. In this example, high cut scores will report a negative trend, while a cut score just above 500 will capture the slice across the distributions with the largest positive vertical trend.

The pogo stick analogy is apt because, even if test score distributions on NAEP and State tests have equal vertical distances at respective percentile levels (an unrealistic assumption), PAC-based score trends reflect only a slice across the test score distributions of interest. The pliability of PAC-based score trends under choices of cut score is not addressed in PAC-based trend comparisons, thus these comparisons are akin to the comparison of the height of children on pogo sticks.

The extent of this pliability in real data is nontrivial. Figure 2 shows the pliability of PAC-based trends from the first fifteen states in alphabetical order for NAEP Grade 4 Reading from 2003 to 2005. For each state shown, three PAC-based trends are calculated from the three NAEP cutpoints: the changes in the percents of students above Basic, Proficient and Advanced. The range of those three trends are shown in the figure. For example, Arizona's PAC-based score trend ranges from +1% to -2% depending on the cut score selected. Connecticut's PAC-based score trend may range from -5% to -1% for cut scores between Basic and Advanced. Space limitations prevent displays across all states, grades, subjects and State testing programs, but these fifteen states are representative of PAC-based trend pliability over the data used in this study.

Insert Figure 2 About Here

Pliability does not cast doubt on tests or test results as a whole. Pliability is statistic-specific – it describes the range of interpretations of a statistic under decisions that may be considered arbitrary. The pliability in this figure demonstrates that attempts to generalize from a single PAC-based trend to the trajectory of a full distribution are, to the degree shown, short-sighted. This short-sightedness is compounded in comparisons of PAC-based trends across testing programs. If a statistic has high pliability, this does not suggest that the data are flawed. Rather, a different statistic or graphical display should be chosen to summarize the data.

Alternative candidates for two-wave trend depiction include average-based statistics, such as effect sizes, and changes-in-percentiles, for example, the difference between medians. Average-based statistics are sensitive to every point in the respective score distribution and are

statistically convenient, but average-based trends are pliable under the choice of score scale. Spencer (1983) describes the conditions under which monotone transformations can reverse the ordering of averages, the sign of trends, and interpretations of trend magnitude. Likewise, comparisons of percentiles are pliable under both the choice of scale transformation and the choice of percentile level. For applications within a single testing program, average-based statistics will be robust and descriptive. However, for comparisons across testing programs, the vagaries of multiple different score scales may be such that a scale-invariant trend statistic is most appropriate. This next section introduces a trend statistic that is remarkably impliable and particularly well suited for the comparison of score trends across tests with different score scales.

A Scale-Invariant Framework for Comparing Test Score Trends

The cornerstone of this scale-invariant framework is the Probability-Probability (PP) plot (Gnanadesikan, 1977). A PP plot can be defined for test score distributions at times 1 and 2 that returns a proportion $p_1 = F_1 [F_2^{-1}(p_2)]$ for all p_2. The PP Plot can be interpreted as returning the proportion of time 1 examinees below given percentiles of the time 2 distribution. Equivalently, given a score, x, the PP Curve contains (p_2, p_1), the proportions from each distribution below that score. Visually the PP Plot is contained within the unit square, and, for continuous distributions, PP curves travel from the origin to the point (1,1). Deviations above the $p_1 = p_2$ diagonal reflect that the proportion of time 1 students below a score is greater than the proportion of time 2 students below that same score. Thus, if the PP curve is largely above the diagonal, this denotes a positive trend. Figure 3 shows a PP plot constructed from the CDFs in Figure 1. PP plots have been used for distributional comparisons in the context of gaps by Haertel, Thrash and Wiley (1978), Spencer (1983), and Livingston (2006). PP plots have conceptual and mathematical ties

to the nonparametric Mann-Whitney *U* statistic, Receiver Operator Characteristic (ROC) Curves, and Lorenz Curves.

Insert Figure 3 About Here

The PP plot is constructed solely from vertical slices across CDFs. A monotone transformation of scale may contort the CDFs horizontally, but will not change the vertical relationships between the cumulative proportions. As such, the PP plot is invariant to monotone transformations of the score scale, and any statistic derived from the PP plot is likewise transformation-invariant. A useful statistic is the area under the PP curve, which is equal to the probability that a randomly drawn test score from time 2 is greater than a randomly drawn test score from time 1. This statistic can be simply designated $P(X_2>X_1)$. In the ROC literature, $P(X_2>X_1)$ is often called the Area Under the ROC Curve (AUC) (e.g., Hanley & McNeil, 1983), though the use and interpretation of this statistic differs in that context. In the nonparametric literature, $P(X_2>X_1)$ can be shown to be the expected value of a linear transformation of the Mann-Whitney *U* statistic, U/mn, where m and n are the numbers of scores in the time 1 and 2 distributions respectively. $P(X_2>X_1)$ lends itself reasonably well to interpretation, where 50% is no overall trend; values greater than 50% suggest an positive trend (time 2 scores are greater than time 1 scores); and values less than 50% suggest a negative trend.

The usefulness of this statistic is that it is invariant to discretionary choices such as cut score, percentile, and score scale. $P(X_2>X_1)$ does not change under any positive monotone transformation. Any transformation of $P(X_2>X_1)$ is therefore also scale-invariant. A second statistic of interest assumes $P(X_2>X_1)$ arises from two normal distributions with unit variance

separated by a distance *V* (Ho & Haertel, 2005). The transformation $V = \sqrt{2} * \Phi^{-1}\left(P(X_2 > X_1)\right)$ returns this statistic (Marzban, 2004). The usefulness of the *V* statistic is that it can be interpreted loosely as a scale-neutral effect size. Unlike the usual effect size *d* (Cohen, 1988), this formulation is impliable under scale transformations. To return to the analogy, these statistics help to remove the pogo sticks from the children to facilitate comparison of their heights.

State testing programs do not report scale-invariant statistics such as $P(X_2>X_1)$ or *V*. However, PP Plots can be estimated by reported data. If the score scales of a test are linked from 2003 to 2005, and cut scores remain the same, any pair of reported PAC statistics defines a point on a PP Plot. As an example, if a state reports that 55% of students are proficient in 2003 and 60% of students are proficient in 2005, then, by definition, 40% of students in 2005 are below the 45th percentile of the 2003 distribution. To put this another way, 40% of time 2 students are below a particular score, and 45% of time 1 students are below that same score. The point (0.4, 0.45) can be plotted on the PP plot, and these points may be used to estimate scale-invariant statistics.

NAEP reports estimates of two hundred 2003-to-2005 state score trends, representing 50 states x 2 Subjects (Reading and Mathematics) x 2 Grades (4 and 8). Of two hundred possible comparable State trends, 82 viable State trends were obtained. The reasons for State trend exclusion are described in Table 1. Some of the most obvious reasons for exclusion are changes in testing programs, score scales, or cut scores between 2003 and 2005. If the definition of proficiency has changed, interpretations of the change in the percent of proficient students as a trend are flawed. Another common reason for State trend exclusion is that states may not test in grade 4 or grade 8 in both years. Though many of these states test and report results for adjacent

grades, e.g., grades 3, 5 or 7, including these results into a State vs. NAEP comparison makes the implicit assumption that trend results from adjacent grades should be similar. This analysis does not make that assumption. Every attempt was made to be conservative about the selection of State score trends. While these precautions result in a smaller sample size, the results show that statistically significant and substantively meaningful conclusions can still be drawn. Data were obtained from state testing websites and reports, and efforts were made to verify the data with state testing representatives from all 50 states.

Insert Table 1 About Here

Table 2 shows the number of PP pairs that were used to estimate scale-invariant State trends. The number of pairs is equal to the number of cut scores for which states reported statistics. This is also equal to the number of proficiency categories minus one. Some states report proportions in only two categories (one cutpoint delineating proficient and not proficient) or in only three categories (two cutpoints delineating basic, proficient and advanced). These states were not included in the analysis, as information about the score distributions remains impoverished at this level. For states with only one PP pair, for example, there is only as much information as in the commonplace and short-sighted PAC-based trends. If assumptions are made about the form of the test score distribution, scale-invariant effect sizes may be readily calculated, but, as Figure 1 helps to show, the shapes of large-scale test score distributions are surprisingly unpredictable over time.

For states with three or more PP pairs, an interpolation procedure was used within the unit square to approximate the PP curve. The theoretical points (0,0) and (1,1) were added to the

state-reported PP points, representing the assumption that there exists an extremely low and an extremely high score that bound the scores of both distributions. These PP points were plotted, and a Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) interpolating function was used to estimate the area under the curve by numerical integration. PCHIP was implemented by the MATLAB program. Simulation studies support the use of this estimation procedure for a range of respectively normal curves where three or more PP pairs are available. In addition, visual inspections of all plots were conducted. As a point of fact, all trend estimates from PP pairs make a kind of parametric assumption in the interpolation method for the PP curve. It is only from the complete distributions that a truly "scale-invariant" measure may be calculated (Macmillan and Creelman, 1996). Nonetheless, every additional PP pair included in the analysis acts to substantially improve the depiction of the full distributions and thus, too, the depiction of the score trend of interest.

Insert Table 2 About Here

Scale-invariant NAEP trends are estimated through a two-stage interpolation procedure. NAEP reports three PAC statistics and five percentile statistics for each score distribution, providing a total of eight points for each CDF. Percentiles are unlikely to align along any given cut score, so PP pairs cannot be obtained from percentiles. Instead, points between the two CDFs are interpolated to obtain a large number of PP pairs for matched cut scores. These PP pairs are located on a PP plot, and interpolation and integration proceeds as before for the PP curve.

Effect size-based trend statistics are generally attenuated by measurement error, but NAEP reports statistics that are corrected for this effect. In contrast, State effect sizes are biased towards zero due to measurement error. If State *V* statistics are treated like traditional effect sizes, they can be corrected by disattenuating them in inverse proportion to the square root of the reliability of the test (Hedges & Olkin, 1985). As reliabilities for State assessments are not always reported, the uncorrected State statistics are used. As the results will show, if the reliabilities of State tests are taken into account, disattenuation will *increase* the degree of average State-NAEP trend discrepancy.

Scale-Invariant Comparisons of State and NAEP Trends

Figure 4 shows the 82 comparable State and NAEP trends using the *V* statistic, which affords an interpretation as a scale-invariant effect size. The *V* statistic is the chosen metric for reporting because the proportion metric of the $P(X_2>X_1)$ statistic is not well suited for aggregation. The null hypothesis of a matched-sample *t*-test for the equality of the mean *V* statistics can be rejected; $t(81)=6.06$; $p<0.001$. The average State trend can be said to be significantly more positive than the average NAEP trend. The centroid is shown in the first quadrant, below the $y=x$ diagonal, as a large gray dot. The average NAEP trend for these 82 state-subject-grade combinations is +0.031 standard deviation units, and the average State trend is +0.115 standard deviation units. Paired trend statistics are shown broken down by quadrant and diagonal. More than 3 out of 4, or 62 of 82 trend pairs (76%) are below the diagonal; these are state-subject-grade combinations where State trends are more positive than NAEP trends. There are 4 instances (5%) where the NAEP trend is positive and the State trend is negative, and there are 21 instances (26%) where the State trend is positive and the NAEP trend is negative.

Slightly fewer than 1 out of 3 instances (30%) show a difference in sign. None of these comparisons takes the measurement error of state tests into account. Disattenuation will spread the points horizontally away from the *y*-axis in approximate inverse proportion to the square root of the reliability of the state tests. Repeating the *t*-test assuming imperfect state test reliabilities would augment the overall finding of significantly different trends.

Insert Figure 4 About Here

Figures 5a-d show the same 82 data points broken down by subject and grade. Figure 5a in the upper left shows that average NAEP and State trends are significantly different for grade 4 Reading; $t(18)=2.99$; $p<0.01$. Figure 5b in the lower left shows that average NAEP and State trends are significantly different in grade 8 Reading; $t(20)=3.97$; $p<0.01$. Figure 5c in the upper right shows that NAEP and State trends are not significantly different in grade 4 Math; $t(18)=1.41$; $p\approx0.177$. Figure 5d in the lower right shows that average NAEP and State trends are significantly different in grade 8 Math; $t(22)=3.83$; $p<0.01$. Grade 4 Math is the only case where NAEP and State average trends are not significantly different.

Both NAEP and State results show that average Math trends are more positive than average Reading trends. However, average NAEP trends are near zero in all cases with the exception of Grade 4 Math, whereas State trends are all positive. These results are consistent with the hypothesis that increased attention to State test content leads to improved performance on State tests but not on NAEP; Grade 4 Math is the exception.

Insert Figures 5a-d About Here

One of the key advantages of the scale-invariant framework is that the magnitudes of trends can be meaningfully compared even when their signs are the same. As an example, Idaho's Mathematics trends shown in Figures 5c and 5d are positive for both NAEP and State tests, but the scale-invariant effect sizes allow for the observation that the State test score trend is substantially more positive than the NAEP trend. Analogous comparisons of PAC-based trends are misleading to the extent that the magnitude of the trend depends critically on the location of the cut score on the distributions. In Figures 4 and 5, the pogo sticks have been removed.

Substantive Issues Arising in State-NAEP Comparisons

The release of the 2005 NAEP results in October of 2005 inspired a number of newspaper articles and reports about State-NAEP discrepancies. State-NAEP comparisons have a long history, and they usually find that State gains are greater than NAEP gains (e.g. Linn, Graue, & Sanders, 1990; Koretz & Barron, 1998; Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz & Hamilton, 2006). However, the scope of recent national educational policies have increased the stakes on State-NAEP comparisons. A substantial amount of recent attention to State-NAEP comparisons concerns discrepancies in the percent of proficient students reported by NAEP and State tests (Linn, Baker, & Betebenner, 2002; Hoover, 2003), but those discrepancies are readily (if not trivially) explained by differences in the results of judgmental standard setting procedures (Linn, 2003; Hambleton, & Pitoniak, 2006). In contrast, trend discrepancies are harder to describe but vastly more relevant to judging the success of educational policies, and the reporting and framing of trend discrepancies is worth discussing here in the context of these scale-invariant results.

As examples, three reports by the Thomas B. Fordham Foundation (2005), the Education Trust (Hall & Kennedy, 2006) and Education Week (2006) compared the changes in the percent of Proficient students as reported by NAEP and State tests. The weaknesses of PAC-based trends are clearly shown by Holland (2002) and Figure 1, but given the broad interpretability and use of PAC-based trends, their use in these reports is predictable. Technical issues aside, these reports occasionally differ in their choice of State trend to report or whether to report a State trend at all. These differences reflect different substantive beliefs about the interpretability of State trends and the comparability of State trends to NAEP.

The Education Week article reports trends for State tests when trend lines and cut scores change, though it adds a footnote to inform readers that results may not be comparable. Both Education Week and the Fordham Foundation use 2004-2005 trends if no 2003 test results are available. All three reports use adjacent grades when grade 4 or 8 comparisons are not available. The first Education Week strategy, footnote aside, results in a number that can at best be ignored and at worst misinterpreted and misused. Reporting a two-year trend instead of a three-year trend or a grade 3 trend instead of a grade 4 trend implies that trends should be commensurate over years and adjacent grades. This paper does not make those assumptions.

By the inclusion and juxtaposition of certain State trends with NAEP trends, this paper does facilitate some of the same questionable comparisons as previous reports. These assumptions include the idea that trends from different seasons should be comparable to Spring trends, as some states (e.g., Indiana, Michigan and Wisconsin) do not test in the Spring. It is clear that the performance of a state that tests in the Fall should be lower than it would if it tested in the Spring, but the juxtaposition of Fall State with Spring NAEP results implies that *gains* in the Fall may be expected to be similar to gains in the Spring. This is a tenuous assumption, as

Fall gains may arguably depend more on teacher-student interactions from previous grade levels, and trends at different or even adjacent grade levels may not be expected to be similar.

Another clearly short-sighted assumption these comparisons encourage is that English and Language Arts (ELA) test score trends are comparable to Reading test score trends, as Reading is usually considered a subset of ELA. For states that report ELA results instead of reporting Reading results (e.g., California, Louisiana, Massachusetts, and South Carolina), content discrepancies are certainly a plausible explanation for State-NAEP discrepancies.

Even when State tests are named, like NAEP tests, "Reading" and "Mathematics," content discrepancies are plausible explanations for State-NAEP discrepancies. Ho (2005) categorized State test items using the NAEP framework to evaluate whether content could account for trend discrepancies. Wei, Lukoff, Shen, Ho and Haertel (2006) conducted a similar analysis on California State tests. Neither found that test content could fully account for trend discrepancies. Notably, they both describe the difficulties in coding State test items using the NAEP framework. This may be taken as a testament to the difficulty of cross-classifying items into test frameworks that are were not originally intended to generate the items. Given a perspective that NAEP and State tests are designed to assess proficiency along different content dimensions, State-NAEP discrepancies are not cause for controversy but a baseline expectation.

Other discrepancy hypotheses exist. Koretz, McCaffrey and Hamilton (2001) present a framework for considering the validity of gains under high-stakes conditions. Trends for a high-stakes, or "focal" test, may differ from trends on a low-stakes test like NAEP (an "audit" test) for a number of reasons, including different "elements of performance" sampled by both tests, different examinee sampling frames, or differing changes in student motivation. Koretz (2005) and Ho and Haertel (2006) have expanded this framework to include still other hypotheses for

trend discrepancies. In this context, each data point in Figure 4 represents the seed for a much larger study that addresses each of these hypotheses in turn. Just as the height of two children may be expected to differ, so too may the results of two very different tests. These differences should be thoughtfully modeled, not cast in terms of a question for which there is only one answer.

Conclusions

Trend comparisons require both technical care and substantive consideration. As useful as PAC statistics have been in communicating test results to the public, their properties as trend statistics render them ill-suited for trend comparison. It seems perfectly reasonable to maintain PAC statistics as a primary means of NCLB reporting while conducting more serious trend analyses using statistics with better properties. While averages and average-based statistics should be the default consideration, the $P(X_2>X_1)$ and V statistics are also strong candidates for trend reporting, especially when dual and different score scales come into play as they do in the arena of State-NAEP trend comparisons.

As NAEP adjusts to its confirmatory role, there must be an active effort to temper expectations that NAEP and State results should be identical. Braun and Mislevy (2005) usefully describe the many intuitions about testing that drive stakeholder expectations of results, including "A test measures what it says at the top of the page," and "A test is a test is a test" (p. 492). As they note, these intuitions may serve stakeholders well for many interpretations and uses of test scores. In high-stakes comparisons of score trends, however, there must be greater attention paid to the vast differences that may exist between the content and format of the two tests, what they are designed to measure, how they are administered, how students engage the test, and how the test results are meant to be used.

References

Braun, H., & Mislevy, R. (2005). Intuitive test theory. *Phi Delta Kappan, 86*(7):489-497.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum, New York, NY.

Education Week (2006). *Quality Counts At 10: A Decade of Standards-Based Education*. Vol. 25, Issue 17.

Fuller, B., Gesicki, K., Kang, E., & Wright, J. (2006). *Is the No Child Left Behind Act Working? The Reliability of How States Track Achievement.* Working Paper 06-1. University of California, Berkeley: Policy Analysis for California Education.

Ferrara, S. (2005). Editorial. *Educational Measurement: Issues and Practice*. *24*(1):1-2.

Ferrara, S. (2006). Editorial. *Educational Measurement: Issues and Practice*. *25*(3):1-3.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. John Wiley & Sons, New York.

Haertel, E., Thrash, W., & Wiley, D. (1978). Metric-free distributional comparisons. Report. ML-Group for Policy Studies in Education, Chicago, IL.

Hall, D., & Kennedy, S. (2006). *Primary Progress, Secondary Challenge*. Report. Downloaded in March, 2006 from http://www2.edtrust.org/NR/rdonlyres/15B22876-20C8-47B8-9AF4-FAB148A225AC/0/PPSCreport.pdf

Hambleton, R.K., & Pitoniak, M.J. (2006). Setting performance standards. In R.L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 531-578). American Council on Education and Praeger Publishers. Westport, Connecticut.

Hanley, J.A., & McNeil, B.J. (1983). A method of comparing the areas under Receiver Operating Characteristic curves derived from the same cases. *Radiology, 178*:839-843.

Hedges, L.V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press Orlando, San Diego.

Ho, A.D. (2005). *Comparing Score Trends on High-Stakes and Low-Stakes Tests Using Metric-Free Statistics and Multidimensional Item Response Models*. Unpublished Doctoral Dissertation. Stanford University.

Ho, A.D., & Haertel, E.H. (2006). Metric-free measures of test score trends and gaps with policy-relevant examples. CSE Technical Report #665. University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST), Los Angeles, CA.

Holland, P. (2002). Two measures of change in the gaps between the CDFs of test-score distributions. *Journal of Educational and Behavioral Statistics*, *27*(1):3–17.

Hoover, H.D. (2003). Some common misconceptions about tests and testing. *Educational Measurement: Issues and Practice*, *22*(1), 5-14.

Klein, S., Hamilton, L., McCaffrey, D., & Stecher, B. (2000). What do test scores in Texas tell us? Report, The RAND Corporation, Santa Monica, CA.

Koretz, D. (2005). Alignment, high stakes, and the inflation of test scores. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement* (104th Yearbook of the National Society of Education, Part 2, pp. 99-118). Malden, MA: Blackwell.

Koretz, D., & Barron, S. (1998). The validity of gains on the Kentucky Instructional Results Information System (KIRIS). Report, The RAND Corporation, Santa Monica, CA.

Koretz, D., & Hamilton, L. (2006). Testing for accountability in K-12. In R.L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 531-578). American Council on Education and Praeger Publishers. Westport, Connecticut.

Koretz, D., McCaffrey, D., & Hamilton, L. (2001). Toward a framework for validating gains under high-stakes conditions. CSE Technical Report #551, University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST), Los Angeles, CA.

Lee, J. (2006) *Tracking Achievement Gaps and Assessing the Impact of NCLB on the Gaps: An In-Depth Look into National and State Reading and Math Outcome Trends*. Civil Rights Project. Harvard University, Cambridge, MA.

Linn, R.L. (2003). Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives, 11*(31). Retrieved March 15, 2007 from http://epaa.asu.edu/epaa/v11n31/

Linn, R.L., Baker, E.L., & Betebenner, D.W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31*(6):3-16.

Linn, R., Graue, M., & Sanders, N. (1990). Comparing state and district results to national norms: The validity of claims that "everyone is above average". *Educational Measurement: Issues and Practice*, *9*(3):5–14.

Livingston, S.A. (2006). Double P-P Plots for Comparing Differences Between Two Groups. *Journal of Educational and Behavioral Statistics*, *31*(4):431-435.

Macmillan, N.A., & Creelman, C.D. (1996). Triangles in ROC space: History and theory of "nonparametric" measures of sensitivity and response bias. *Psychonomic Bulletin and Review*. *3*(2):164-170.

Marzban, C. (2004). A comment on the ROC curve and the area under it as performance measures. Downloaded in June, 2005 from http://www.csee.wvu.edu/~ross/courses/sp06/biom693/reading/MarzbanROC_2004.pdf

Spencer, B. (1983). On interpreting test scores as social indicators: Statistical considerations. *Journal of Educational Measurement*, *20*:317–333.

Thomas B. Fordham Foundation (2005). *Gains on State Reading Tests Evaporate on 2005 NAEP*. Report. Downloaded in November, 2005 from http://www.edexcellence.net/foundation/about/press_release.cfm?id=19

Wei, X., Lukoff, B., Shen, X., Ho, A.D., & Haertel, E.H. (2006). Using test content to address trend discrepancies between NAEP and California state tests. Paper to be presented at the 2006 meeting of the American Educational Research Association. San Francisco, CA.

Figure 1. The Pliability of PAC-Based Trends (Dotted) Under Selections of Cut Score

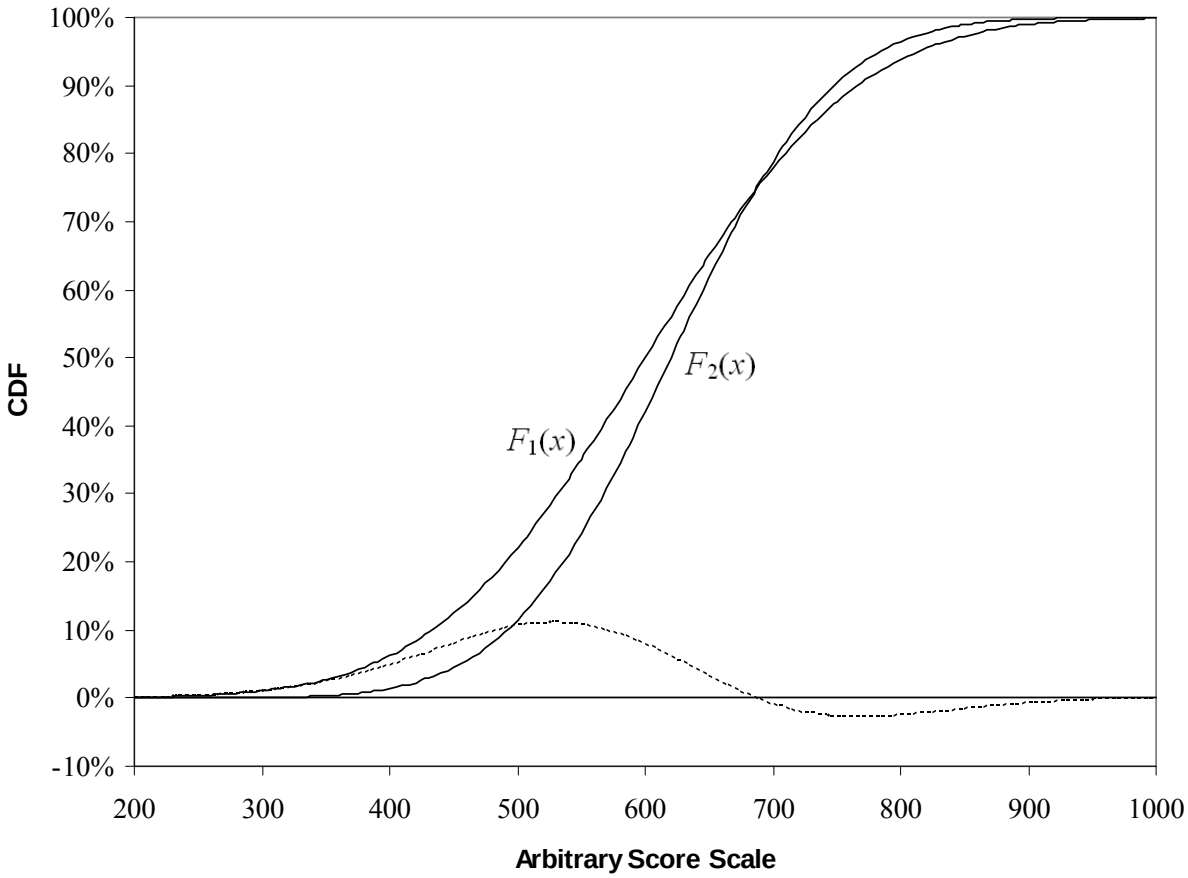

Figure 2: NAEP Reading Grade 4, Pliability of PAC-Based Trends for the First 15 Alphabetized States Under Three Cut Scores: (b)asic, (p)roficient, and (a)dvanced.

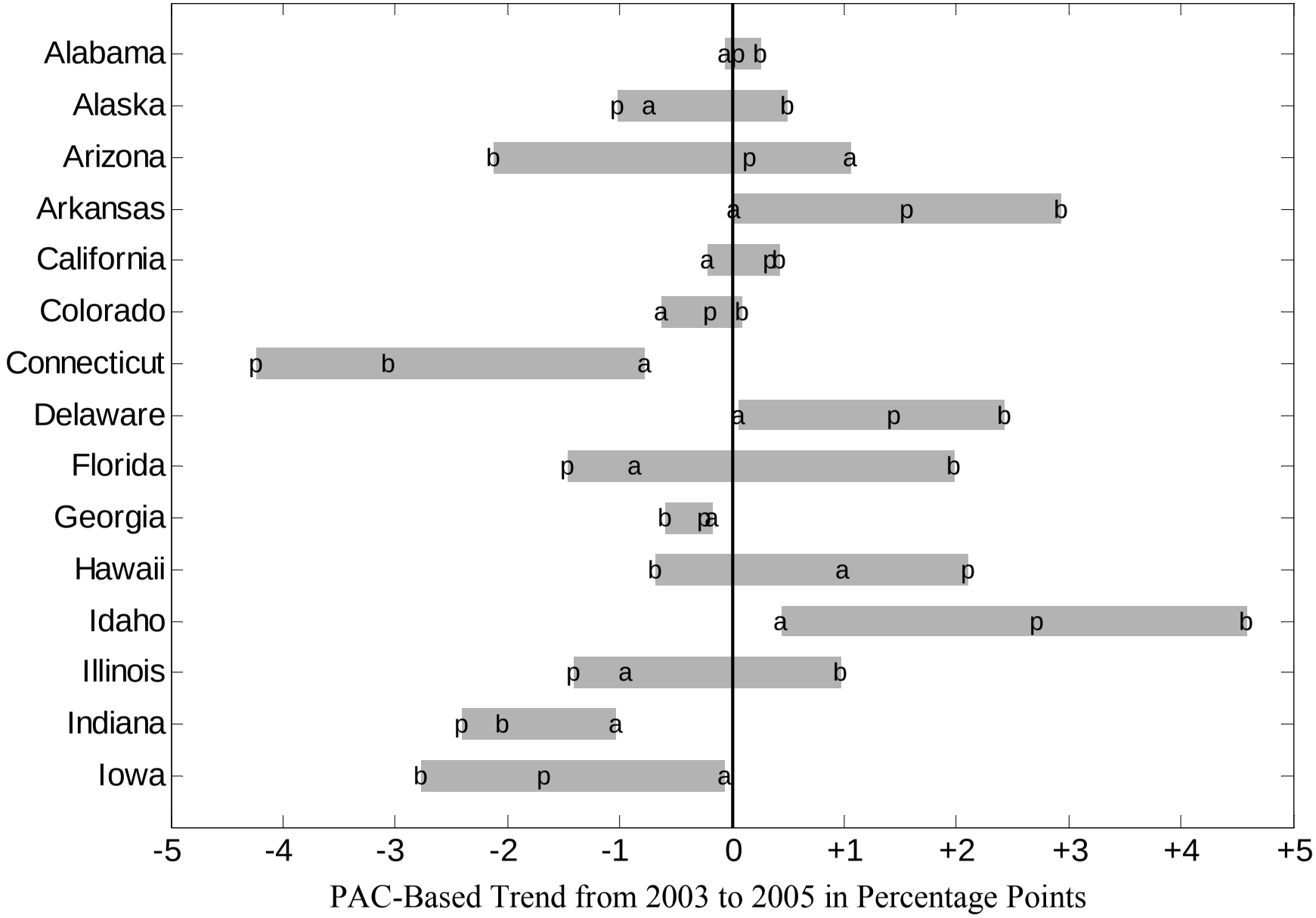

Figure 3. A Probability-Probability (PP) Plot Generated from the CDFs from Figure 1. The Diagonal (Dotted Line) is Shown for Reference.

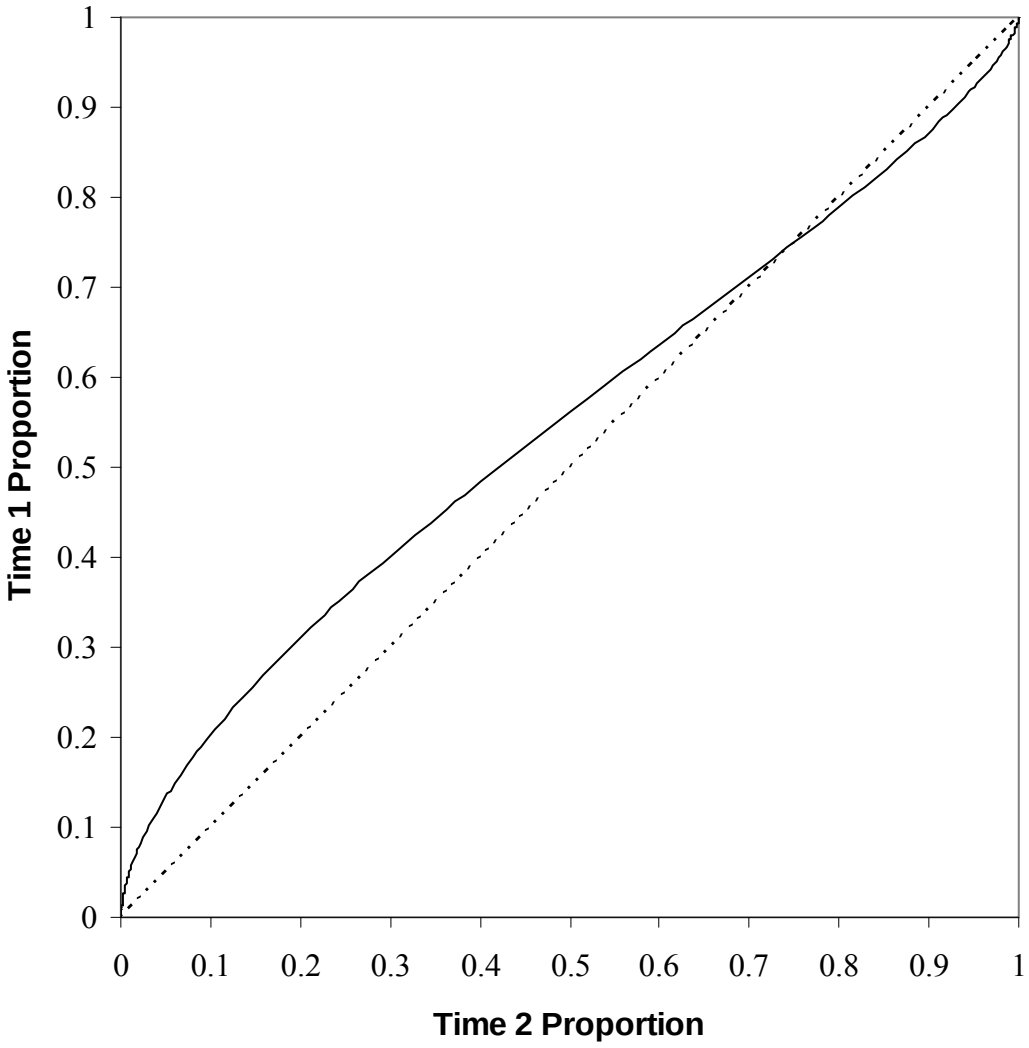

Table 1: Reasons for State Trend Exclusion by Subject and Grade

Subject:	Reading		Mathematics	
State Grade:	Grade 4	Grade 8	Grade 4	Grade 8
Alabama	No '03	No '03	No '03	No '03
Alaska	New Cuts	New Cuts	New Cuts	New Cuts
Arizona	New Cuts	New Cuts	New Cuts	New Cuts
Arkansas	New Trend	New Trend	New Trend	New Trend
California				Two Tests
Colorado			No '03	
Connecticut				
Delaware	No Data		No Data	
Florida				
Georgia	<3 Cutpoints	<3 Cutpoints	<3 Cutpoints	<3 Cutpoints
Hawaii	No '03		No '03	
Idaho				
Illinois	No Data		No Data	
Indiana	No Fall '02	<3 Cutpoints	No Fall '02	<3 Cutpoints
Iowa	<3 Cutpoints	<3 Cutpoints	<3 Cutpoints	<3 Cutpoints
Kansas	No Data			No Data
Kentucky		No Data	No Data	
Louisiana				
Maine				
Maryland	No '03	<3 Cutpoints	No '03	<3 Cutpoints
Massachusetts		No Data		
Michigan		No Winter Test		
Minnesota	No Data	No Data	No Data	No Data
Mississippi				
Missouri	No Data	No Data		
Montana				
Nebraska	No '03	No '03	No '03	No '03
Nevada	No Data	No '03	No Data	No '03
New Hampshire	No Data	No Data	No Data	No Data
New Jersey	<3 Cutpoints	No '03	<3 Cutpoints	No '03
New Mexico	New Test	New Test	New Test	New Test
New York				
North Carolina				
North Dakota	New Cuts	New Cuts	New Cuts	New Cuts
Ohio	New Cuts	New Cuts	New Cuts	New Cuts
Oklahoma	New Test		New Test	
Oregon	No Data	<3 Cutpoints	No Data	<3 Cutpoints
Pennsylvania	No Data		No Data	
Rhode Island	Unreleased	Unreleased	Unreleased	Unreleased
South Carolina				
South Dakota	New Cuts	New Cuts	New Cuts	New Cuts
Tennessee	<3 Cutpoints	<3 Cutpoints	<3 Cutpoints	<3 Cutpoints
Texas				
Utah	New Cuts	New Cuts	New Cuts	New Cuts
Vermont	Unreleased	Unreleased	Unreleased	Unreleased
Virginia	No Data	<3 Cutpoints	No Data	<3 Cutpoints
Washington		No Data		No Data
West Virginia	No 2003	No 2003	No 2003	No 2003
Wisconsin				
Wyoming				

Table 2: Number of Cutpoints Used to Estimate State PP Curves by State, Subject and Grade.

Subject:	Reading		Mathematics	
State Grade:	Grade 4	Grade 8	Grade 4	Grade 8
Alabama	-	-	-	-
Alaska	-	-	-	-
Arizona	-	-	-	-
Arkansas	-	-	-	-
California	4	4	4	-
Colorado	3	3	-	3
Connecticut	4	4	4	4
Delaware	-	4	-	4
Florida	4	4	4	4
Georgia	-	-	-	-
Hawaii	-	3	-	3
Idaho	3	3	3	3
Illinois	-	3	-	3
Indiana	-	-	-	-
Iowa	-	-	-	-
Kansas	-	4	4	-
Kentucky	7	-	-	7
Louisiana	4	4	4	4
Maine	3	3	3	3
Maryland	-	-	-	-
Massachusetts	3	-	3	3
Michigan	3	-	3	3
Minnesota	-	-	-	-
Mississippi	3	3	3	3
Missouri	-	-	4	4
Montana	3	3	3	3
Nebraska	-	-	-	-
Nevada	-	-	-	-
New Hampshire	-	-	-	-
New Jersey	-	-	-	-
New Mexico	-	-	-	-
New York	3	3	3	3
North Carolina	3	3	3	3
North Dakota	-	-	-	-
Ohio	-	-	-	-
Oklahoma	-	3	-	3
Oregon	-	-	-	-
Pennsylvania	-	3	-	3
Rhode Island	-	-	-	-
South Carolina	3	3	3	3
South Dakota	-	-	-	-
Tennessee	-	-	-	-
Texas	4	4	4	4
Utah	-	-	-	-
Vermont	-	-	-	-
Virginia	-	-	-	-
Washington	3	-	3	-
West Virginia	-	-	-	-
Wisconsin	3	3	3	3
Wyoming	3	3	3	3

Figure 4: NAEP vs. State Score Trend Discrepancies; All 82 State-Subject-Grade Combinations.

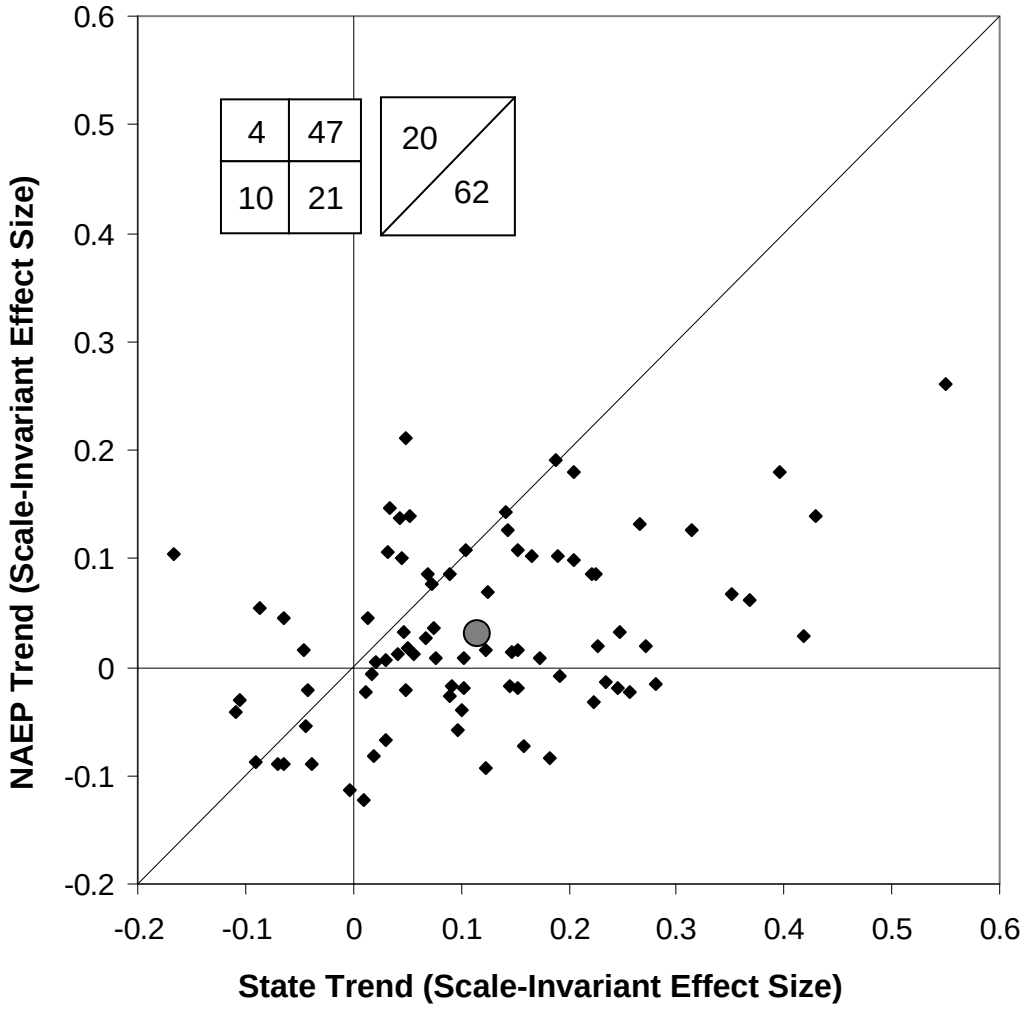

Figure 5a-d: Comparing NAEP vs. State Score Trend Discrepancies using *V* Statistics, a) Reading Grade 4, b) Reading Grade 8, c) Mathematics Grade 4, d) Mathematics Grade 8.

a c R4 M4

b d R8 M8